Available online at www.sciencedirect.com
ScienceDirect
Journal of Retailing and Consumer Services 14 (2007) 359–368
JOURNAL OF RETAILING AND CONSUMER SERVICES
www.elsevier.com/locate/jretconser

Latent class modeling of website users' search patterns: Implications for online market segmentation

José G. Dias[a,*], Jeroen K. Vermunt[b]

[a]Department of Quantitative Methods and UNIDE, ISCTE, Higher Institute of Social Sciences and Business Studies, Edifício ISCTE, Av. das Forças Armadas, 1649–026 Lisboa, Portugal
[b]Department of Methodology and Statistics, Tilburg University, P.O. Box 90153 NL-5000 LE Tilburg, The Netherlands

Abstract

Appropriate modeling of web use patterns may yield very relevant marketing and retailing information. We propose using a model-based clustering approach for market segmentation based on website users' search patterns. We not only provide a detailed discussion of technical issues such as the problem of the selection of the number of segments, but also a very interesting empirical illustration of the potentials of the proposed approach.

Keywords: Web usage mining; Market segmentation; Latent class model

1. Introduction

In recent years web usage mining has become a very important topic of research in the field of data mining. Web usage mining involves using data mining techniques in the discovery of web navigation patterns from web log data, also referred to as click stream data (Hand et al., 2001; Spiliopoulou and Pohle, 2001; Huang et al., 2006). The idea is that the analysis of the sequence of web pages requested by users within a particular web site may provide a better understanding and prediction of users' behavior, and may thus be used to improve the design of the web site concerned. For example, web mining of online stores may yield information on the effectiveness of marketing and web merchandizing efforts, such as how the consumers start the search, which products they see, and which products they buy (Shahabi et al., 1997; Lee et al., 2001). It has, however, been shown that traditional data mining algorithms may not be suited for the discovery of web usage patterns (Spiliopoulou and Pohle, 2001).

An approach that has been used extensively for web mining views a particular web user's navigation pattern on a web site as a Markovian process (Sarukkai, 2000; Dongshan and Junyi, 2002). A Markov model is built for predicting the next web page requested by a web user. A limitation of simple Markov models is that they do not take into account differences in user preferences.

In this paper we propose capturing unobserved heterogeneity among web users using a model with a discrete latent variable. This latent variable allows the segmentation of the data set into clusters and, once the model is learned, web users can be assigned into clusters. More specifically, we use a model known as finite mixture of Markov chains, which has been applied as a model-based clustering tool for classifying web users into different categories (Cadez et al., 2003; Sen and Hansen, 2003). This model has been referred to as mixed Markov model in applied literature (Poulsen, 1990; van de Pol and Langeheine, 1990). Despite of the widespread application of finite mixture models, the number of latent segments to retain is still a very important topic of research. Whereas most researchers use AIC and BIC for determining the dimension of these models, there is evidence that the new AIC3 measure is more appropriate for discrete data (Andrews and Currim, 2003).

*Corresponding author. Tel.: +351 217 903 228; fax: +351 217 903 004.
E-mail addresses: jose.dias@iscte.pt (J.G. Dias), J.K.Vermunt@uvt.nl (J.K. Vermunt).

0969-6989/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jretconser.2007.02.007

The goal of this paper is fifth fold. First, it brings together web mining and market segmentation issues in retailing research. Second, we discuss the modeling of web usage patterns by finite mixtures. Third, we discuss the setting of the number of segments and provide a Monte Carlo (MC) study. Fourth, we introduce the Markov map that allows a fast understanding of the dynamics within each segment. Finally, an empirical application shows the relevance of the connection between web mining and marketing segmentation developed here.

Section 2 gives a short review of market segmentation issues in web mining applications. Section 3 presents model specification and estimation of the finite mixture model for sequential data. It also discusses the selection of the number of latent segments using information criteria. Section 4 presents the results from the MC study that was performed to gain more insight in the performance of different model selection criteria. Section 5 illustrates the implications for web usage mining of the theoretical results with a real data set. The paper concludes with a summary of main findings, implications, and suggestions for further research.

2. Market segmentation issues in web mining research

Market segmentation has become a key concept in marketing theory and practice (Wind, 1978; Wedel and Kamakura, 2000). Smith (1956) defined it as: "market segmentation consists of viewing a heterogeneous market as a number of smaller homogeneous markets in response to differing product preferences, (...) attributable to the desires of consumers or users for more precise satisfaction of their varying wants" (p. 6). Viewing a heterogeneous population as being composed of homogeneous subgroups or segments, each of which responds differently to the marketing mix has been shown to greatly increase marketing efficiency (Wedel and Kamakura, 2000).

Segmentation studies to identify those homogeneous groups have two major components: the information used as input, called the 'bases' of segmentation, and the methods used to identify segments/subpopulations based on the input data (Wind, 1978; Wedel and Kamakura, 2000).

Traditionally, segmentation bases have been classified into four categories: demographic variables (e.g., age, sex, household size), geographic variables (e.g., ZIP code, region), psychographic variables (e.g., attitudes, values, lifestyles), and behavioral variables (e.g., frequency of use, usage level). The most effective segmentation strategy is that which best captures differences in the behavior of target subpopulations, for instance, behavior status or behavioral dynamics.

Segmentation methods can be classified into (Wedel and Kamakura, 2000): (i) *a priori* approaches, when the type and number of subpopulations are defined in advance based on specific criteria, usually demographic variables; and (ii) *post hoc* approaches, when the type and number of subpopulations emerge from the segmentation procedure applied to the data after they have been collected. *Post hoc* approaches to segmentation, and in particular those based on finite mixture models have received much attention in the marketing literature (Wedel and Kamakura, 2000). Finite mixture models have been shown to outperform traditional *post hoc* approaches involving cluster analysis (Vriens et al., 1996).

It is therefore not surprising that also in web mining a substantial effort has been put in an attempt to discover groups of users exhibiting similar browsing patterns (Vakali et al., 2004). Whereas the purpose of user clustering is to establish groups of users that present similar browsing patterns, sometimes the aim may be to discover groups of pages with a similar content (Shahabi et al., 1997; Petridou et al., 2006). Poblete and Baeza-Yates (2006) adopted such a supply-oriented viewpoint based on the clustering of pages of web sites. Various examples exist of studies adopting a consumer-oriented viewpoint—which is also our focus—and that involve clustering users. Petridou et al. (2006) proposed clustering web users using a K-means algorithm based on the KL-divergence which measures the "distance" between individual data distributions. A similar approach is adopted by Yang and Padmanabhan (2005), who looked at the number of times a given user visited a given webpage. Smith and Ng (2003) suggested using self-organizing maps (SOMs) of user navigation patterns. However, none of these approaches accounts for the sequential structure of browsing data. This means that consecutive states in a sequence are, in fact, treated as independent observations conditional on cluster membership, an assumption that is rather unrealistic.

Because finite mixtures are probabilistic or stochastic models, it is possible to parameterize the model in ways that respect the structure of data (e.g., taking into account serial dependence) and test for response parameters and the number of subpopulations with appropriate statistical methods (McLachlan and Peel, 2000). From an unsupervised learning perspective, Markov models with latent variables have been an important paradigm for modeling sequential data (Saul and Jordan, 1998; Cadez et al., 2003; Pallis et al., 2005).

3. The latent segment Markov chain model

This section introduces a flexible model for web users' search patterns that accommodates for both heterogeneity and serial dependencies. Consider a sample of n web users. A web user will be denoted by i ($i = 1, \ldots, n$). Each web user is characterized by a sequence of states $\mathbf{x}_i$. Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote a sample of size n. Note that the state of the website user is, in fact, the category to which the current web page belongs to (frontpage, new, etc.). A next state is generated by moving to another new page, where the next state will be the same as the previous if the new pages concerned belong to the same category. A sequence is formed by a series of visited pages.